ПРИЛОЖЕНИЕ
к статье А.Ю. Пронозина, Е.А. Салиной, Д.А. Афонникова
«GBS-DP: биоинформатический конвейер для обработки данных, полученных генотипированием путем секвенирования»

Временные затраты для обработки данных на различных этапах выполнения конвейера GBS-DP для разного количества библиотек ячменя

Этап обработки данных	Входные данные	Выходные файлы	Число библиотек					
			10	50	100	150	200	272
Этап 1: Предобработка данных								
Удаление адаптеров	Сырые прочтения, FASTQ.GZ	Очищенные прочтения, FASTQ	00:02	00:04	00:13	00:20	00:26	00:36
Построение индекса референсного генома	Референсный геном, FASTA	Индекс референсного генома, FAST.idx	00:05	00:05	00:05	00:05	00:05	00:05
Этап 2: Поиск полиморфизмов								
Картирование на референсный геном	Референсный геном, FASTA Предобработанные прочтения, FASTQ	Картированные прочтения, SAM	00:05	00:33	01:15	1:52	2:30	03:24
Сортировка картированных прочтений	Картированные прочтения, SAM	Отсортированные прочтения, BAM Индекс для каждой библиотеки, BAI	>00:01	>00:01	00:01	00:02	00:02	00:02
Поиск одно-нуклеотидных полиморфизмов	Отсортированные прочтения, BAM Индекс для каждой библиотеки, SORT	Результаты поиска полиморфизмов, VCF	00:08	00:36	01:20	1:54	2:28	05:20
Этап 3: Анализ генетического разнообразия								
Подготовка данных для анализа	Результаты поиска полиморфизмов, VCF	Общий файл, содержащий данные о полиморфизмах для всех библиотек, VCF Общий файл, содержащий данные о полиморфизмах для всех библиотек, GDS Или Файлы для каждой хромосомы, содержащие данные о полиморфизмах для всех библиотек, VCF Общий файл, содержащий данные о полиморфизмах для всех библиотек, GDS	00:18	01:39	4:20	8:12	2:54	06:18
Построение филогенетического дерева и кластеризация	Общий файл, содержащий данные о полиморфизмах для всех библиотек, GDS	Филогенетическое дерево, TREE Кластеризация генотипов, PNG	00:03	00:15	00:40	1:10	01:48	04:20
Общее время			00:37	03:12	7:54	13:35	7:19	20:05

Примечание. Время обработки указано в формате часы:минуты. Использовались библиотеки коротких прочтений из проекта PRJEB39633 БД ENA. Для расчетов использовался вычислительный узел кластера ЦКП «Биоинформатика» с процессором AMD EPYC 74521, 32 вычислительными ядрами и объемом оперативной памяти 1 Тб. Для вычислений было использовано 100 Гб оперативной памяти и 20 вычислительных ядер.